Understanding Protein Function on a Genome-scale through the Analysis of Molecular Networks

Mark B Gerstein
Yale

slides at Lectures.GersteinLab.org

(See Last Slide for References & More Info.)
The problem: Grappling with Function on a Genome Scale?

- 250 of ~530 originally characterized on chr. 22
  [Dunham et al. Nature (1999)]

- >25K Proteins in Entire Human Genome
  (with alt. splicing)
Traditional single molecule way to integrate evidence & describe function

Descriptive Name: Elongation Factor 2

Lots of references to papers

Summary sentence describing function:
This protein promotes the GTP-dependent translocation of the nascent protein chain from the A-site to the P-site of the ribosome.

EF2_YEAST
Some obvious issues in scaling single molecule definition to a genomic scale

• Fundamental complexities
  ◇ Often >2 proteins/function
  ◇ Multi-functionality:
    2 functions/protein
  ◇ Role Conflation:
    molecular, cellular, phenotypic
Some obvious issues in scaling single molecule definition to a genomic scale

- Fundamental complexities
  ◦ Often >2 proteins/function
  ◦ Multi-functionality:
    2 functions/protein
  ◦ Role Conflation:
    molecular, cellular, phenotypic

- Fun terms… but do they scale?....
  ◦ Starry night (P Adler, ’94)
Hierarchies & DAGs of controlled-vocab terms but still have issues...

MIPS (Mewes et al.)

GO (Ashburner et al.)

[Seringhaus & Gerstein, Am. Sci. '08]
Towards Developing Standardized Descriptions of Function

- Subjecting each gene to standardized expt. and cataloging effect
  - KOs of each gene in a variety of std. conditions => phenotypes
  - Std. binding expts for each gene (e.g. prot. chip)

- Function as a vector

<table>
<thead>
<tr>
<th></th>
<th>nucleic acids</th>
<th>small molecules</th>
<th>proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>RNA</td>
<td>ATP</td>
<td>Metal</td>
</tr>
<tr>
<td>protein 1</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>protein 2</td>
<td>0</td>
<td>0.9</td>
<td>0</td>
</tr>
<tr>
<td>protein 3</td>
<td>1.0</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>protein 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>protein 5</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>protein 6</td>
<td>0.9</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>protein 7</td>
<td>0</td>
<td>0.8</td>
<td></td>
</tr>
</tbody>
</table>

Interaction Vectors [Lan et al, IEEE 90:1848]
Networks (Old & New)

**Classical KEGG pathway**

- **Fringe:** Vital in boundary formation in developing fly wing.
- **Numb:** mutations impair sensory organs in flies
- **Notch:** with defects, flies develop notches in wings

**Same Genes in High-throughput Network**

- **Itch:** linked to itchy skin in mice
- **DLK1:**
- **Fringe:**
- **Numb:**
- **Notch:**

[Seringhaus & Gerstein, Am. Sci. '08]
Networks occupy a midway point in terms of level of understanding

1D: Complete Genetic Partslist

~2D: Bio-molecular Network Wiring Diagram

3D: Detailed structural understanding of cellular machinery

Networks as a universal language

- Internet [Burch & Cheswick]
- Food Web [Krebs]
- Electronic Circuit [Barabasi]
- Disease Spread [Krebs]
- Protein Interactions [Barabasi]
- Neural Network [Cajal]
- Social Network
Using the position in networks to describe function

Guilt by association

Finding the causal regulator (the "Blame Game")

[NY Times, 2-Oct-05, 9-Dec-08]
Combining networks forms an ideal way of integrating diverse information

- **Metabolic pathway**
- **Transcriptional regulatory network**
- **Physical protein-protein Interaction**
- **Co-expression Relationship**

Genetic interaction (synthetic lethal)
Signaling pathways

Part of the TCA cycle
Outline: Molecular Networks

• Why Networks?
• Predicting Networks (yeast)
  ◊ Propagating known information
• Dynamics & Variation of Networks
  ◊ Across cellular states (yeast)
  ◊ Across environments (in prokaryotes)
• Protein Networks & Variation (yeast & humans)
Example: yeast PPI network

Actual size:
- ~6,000 nodes
  - Computational cost: ~18M pairs
- Estimated ~15,000 edges
  - Sparseness: 0.08% of all pairs
  - (Yu et al., 2008)

Known interactions:
- Small-scale experiments: accurate but few
  - Overfitting: ~5,000 in BioGRID, involving ~2,300 proteins
- Large-scale experiments: abundant but noisy
  - Noise: false +ve/-ve for yeast two-hybrid data up to 45% and 90% (Huang et al., 2007)
Types of Networks

Interaction networks

Nodes: proteins or genes
Edges: interactions

[Horak, et al, Genes & Development, 16:3017-3033]
[DeRisi, Iyer, and Brown, Science, 278:680-686]
[Jeong et al, Nature, 41:411]

Regulatory networks

Metabolic networks
Predicting Networks

How do we construct large molecular networks? From extrapolating correlations between functional genomics data with fairly small sets of known interactions, making best use of the known training data.
Network prediction: known information

Known interactions: 1-2, 2-3
Known non-interactions: 3-4
Unknown: 1-4, 2-3
Network prediction: features

- Example 1: gene expression

\[ x_1 = (0.2, 2.4, 1.5, \ldots) \]
\[ x_2 = (0.8, 2.2, 1.5, \ldots) \]
\[ x_3 = (4.3, 0.1, 7.5, \ldots) \]

\[ \ldots \]
\[ \text{sim}(x_1, x_2) = 0.62 \]
\[ \text{sim}(x_1, x_3) = -0.58 \]

\[ \ldots \]

Gasch et al., 2000
Network prediction: features

• Example 2: sub-cellular localization

\[
x_1 = (1, 1, 0, 0, \ldots) \\
x_2 = (1, 1, 1, 0, \ldots) \\
x_3 = (1, 0, 1, 0, \ldots) \\
\ldots
\sim(x_1, x_2) = 0.81 \\
\sim(x_1, x_3) = 0.12 \\
\ldots
\]
Network prediction: data integration
Learning methods

An endless list:

• Docking (e.g. Schoichet and Kuntz 1991)
• Evolutionary (e.g. Ramani and Marcotte, 2003)
• Topological (e.g. Yu et al., 2006)
• Bayesian (e.g. Jansen et al., 2003)
• Kernel methods
  ◊ Global modeling:
    • em (Tsuda et al., 2003)
    • kCCA (Yamanishi et al., 2004)
    • kML (Vert and Yamanishi, 2005)
    • Pairwise kernel (Pkernel) (Ben-Hur and Noble, 2005)
  ◊ Local modeling:
    • Local modeling (Bleakley et al., 2007)

Let’s compare in a public challenge!
(DREAM: Dialogue for Reverse Engineering Assessment and Methods)
DREAM3: *in silico* regulatory network reconstruction

Actual network → Expression data → Modeling → Predictions

Deletion strains

Noise models

Expression rate models

\[
\text{Prob(signal|point)} = 2\Phi((\text{point} - \text{ref}) / s) - 1
\]

\[
\frac{dy_j}{dt} = a_{j1} - a_{j2}y_j + \sum_{k} b_{jk}y_k
\]

\[
\frac{dy_j}{dt} = \frac{b_{jk}}{1 + \exp\left(a_{jk} + \sum_{l} d_{jl}y_l\right)} - b_{j2}y_j
\]

\[
\frac{dy_j}{dt} = a_{jk} \prod_{i} \left( \frac{b_{ik}}{y_i + b_{ik}} \right) \prod_{l} \left( \frac{y_l^{\psi_{il}}}{y_l^{\psi_{il}} + b_{il}} \right) - a_{j1}y_j
\]

Accuracy (AUC)

<table>
<thead>
<tr>
<th></th>
<th>E. Coli 1</th>
<th>E. Coli 2</th>
<th>Yeast 1</th>
<th>Yeast 2</th>
<th>Yeast 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size-10</td>
<td>0.928</td>
<td>0.912</td>
<td>0.949</td>
<td>0.747</td>
<td>0.714</td>
</tr>
<tr>
<td>Size-50</td>
<td>0.930</td>
<td>0.924</td>
<td>0.917</td>
<td>0.792</td>
<td>0.805</td>
</tr>
<tr>
<td>Size-100</td>
<td>0.948</td>
<td>0.960</td>
<td>0.915</td>
<td>0.856</td>
<td>0.783</td>
</tr>
</tbody>
</table>

[Yip et al., DREAM3]
Our work: efficiently propagating known information

Training set expansion
• Motivation: lack of training examples
• Expand training sets horizontally

Multi-level learning
• Motivation: hierarchical nature of interaction
• Expand training sets vertically

DREAM3 *in silico* regulatory network reconstruction challenge
Global vs. local modeling

Global modeling: build one model for the whole network

Example - Pairwise kernel: consider object pairs instead of individual objects

Problem: $O(n^2)$ instances, $O(n^4)$ kernel elements
Global vs. local modeling

Local modeling: build one model for each node

Model for node 3:

Problem: insufficient and unevenly distributed training data (what if node 3 has no known interactions at all?)
Prediction propagation

• Goal: keep the flexibility of local modeling, but tackle the data sparsity problem
• Motivation: some objects have more examples than others
• Our approach:
  ◊ Learn models for objects with more examples first
  ◊ Propagate the most confident predictions as auxiliary examples of other objects

[Yip and Gerstein, Bioinformatics ('09)]
Observations:

- Highest accuracy by training set expansion
- Over fitting of local modeling without training set expansion
- Prediction propagation theoretically related to co-training (Blum and Mitchell, 1998)
  ◇ Semi-supervised (Similarity with PSI-BLAST)

[Yip and Gerstein, Bioinformatics ('09)]
From horizontal to vertical

Training set expansion
• Motivation: lack of training examples
• Expand training sets horizontally

Multi-level learning
• Motivation: hierarchical nature of interaction
• Expand training sets vertically
Protein interaction

Yeast NADP-dependent alcohol dehydrogenase 6 (PDB: 1piw)

Protein-level features for interaction prediction: functional genomic information

[Yip and Gerstein, in revision]
Domain interaction

Pfam domains: PF00107 (inner) and PF08240 (outer)

Domain-level features for interaction prediction: evolutionary information

[Yip and Gerstein, in revision]
Residue interaction

Interacting residues: 283 (yellow) with 287 (cyan), and 285 (purple) with 285

Residue-level features for interaction prediction: physical-chemical information

[Yip and Gerstein, in revision]
Combining the three problems

- Protein interactions
- Domain interactions
- Residue interactions

i. Independent levels
ii. Unidirectional flow
iii. Bidirectional flow

[Yip and Gerstein, in revision]
Empirical results (AUCs)

<table>
<thead>
<tr>
<th></th>
<th>Ind. levels</th>
<th>Unidirectional flow</th>
<th>Bidirectional flow</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level</strong></td>
<td></td>
<td>PD</td>
<td>PR</td>
</tr>
<tr>
<td>Proteins</td>
<td>71.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Domains</td>
<td>53.18</td>
<td>61.51</td>
<td></td>
</tr>
<tr>
<td>Residues</td>
<td>57.36</td>
<td>54.89</td>
<td>53.81</td>
</tr>
</tbody>
</table>

- Highest accuracy by bidirectional flow
- Additive effect: 2 vs. 3 levels

[Yip and Gerstein, in revision]
Network Dynamics #1: Cellular States

How do networks change across different cellular states?
How can this be used to assign function to a protein?
Dynamic Yeast TF network

- Analyzed network as a static entity
- But network is *dynamic*
  - Different sections of the network are active under different cellular conditions
- Integrate gene expression data

*Luscombe et al. Nature 431: 308*
Gene expression data for five cellular conditions in yeast

<table>
<thead>
<tr>
<th>Cellular condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell cycle</td>
</tr>
<tr>
<td>Sporulation</td>
</tr>
<tr>
<td>Diauxic shift</td>
</tr>
<tr>
<td>DNA damage</td>
</tr>
<tr>
<td>Stress response</td>
</tr>
</tbody>
</table>

[Multi-stage]  [Binary]

[Brown, Botstein, Davis,...]
Backtracking to find active sub-network

- Define differentially expressed genes
- Identify TFs that regulate these genes
- Identify further TFs that regulate these TFs

Active regulatory sub-network
Network usage under different conditions

static

Luscombe et al. Nature 431: 308
Network usage under different conditions

cell cycle
Network usage under different conditions

sporulation
Network usage under different conditions

diauxic shift
Network usage under different conditions

DNA damage
Network usage under different conditions

stress response
Network usage under different conditions

Cell cycle  Sporulation  Diauxic shift  DNA damage  Stress

SANDY:
1. Standard graph-theoretic statistics:
   - Global topological measures
   - Local network motifs

2. Newly derived follow-on statistics:
   - Hub usage
   - Interaction rewiring

3. Statistical validation of results

Luscombe et al. Nature 431: 308
Network usage under different conditions

- Cell cycle
- Sporulation
- Diauxic shift
- DNA damage
- Stress

**SANDY:**
1. Standard graph-theoretic statistics:
   - Global topological measures
   - Local network motifs

2. Newly derived follow-on statistics:
   - Hub usage
   - Interaction rewiring

3. Statistical validation of results
Global topological measures

Indicate the gross topological structure of the network

Degree ($K$) 5
Path length ($L$) 2
Clustering coefficient ($C$) 1/6

Interaction and expression networks are **undirected**

[Barabasi]
Global topological measures for directed networks

Regulatory and metabolic networks are directed
Scale-free networks

Hubs dictate the structure of the network

[Barabasi]
### Analysis of condition-specific subnetworks in terms of global topological statistics

<table>
<thead>
<tr>
<th>Measure</th>
<th>Cell cycle</th>
<th>Sporulation</th>
<th>Diauxic shift</th>
<th>DNA damage</th>
<th>Stress response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outdegree</td>
<td>7.9</td>
<td>6.5</td>
<td><strong>17.1</strong></td>
<td>15.0</td>
<td>9.0</td>
</tr>
<tr>
<td>Indegree</td>
<td>2.0</td>
<td>1.9</td>
<td>1.6</td>
<td>1.6</td>
<td>1.6</td>
</tr>
<tr>
<td>Pathlength</td>
<td>4.5</td>
<td>3.4</td>
<td><strong>2.1</strong></td>
<td>2.0</td>
<td>2.2</td>
</tr>
<tr>
<td>Clustering coeff</td>
<td>0.15</td>
<td>0.14</td>
<td><strong>0.09</strong></td>
<td>0.09</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Multi-stage controllable ticking over of genes at different stages

Binary:
- Quick, large-scale turnover of genes

Luscombe et al. Nature 431: 308
**Summary**

**Cell cycle**  
- multi-stage conditions  
  - less pronounced  
  - longer  
  - more  
  - complex loops (FFLs)

**Sporulation**

**Diauxic shift**  
- binary conditions  
  - more pronounced  
  - shorter  
  - less  
  - simpler (SIMs)

**DNA damage**

**Stress**
Transient Hubs

- Questions:
  ◊ Do hubs stay the same or do they change over between conditions?
  ◊ Do different TFs become important?

- Our Expectations
  ◊ Literature:
    - Hubs are permanent features of the network regardless of condition
  ◊ Random networks (sampled from complete regulatory network)
    - Random networks converge on same TFs
    - 76-97% overlap in TFs classified as hubs (ie hubs are permanent)

Luscombe et al. Nature 431: 308
- Some permanent hubs
  ◊ house-keeping functions

- Most are transient hubs
  ◊ Different TFs become key regulators in the network

- Implications for condition-dependent vulnerability of network

Luscombe et al. Nature 431: 308
transient hubs
permanent hubs
Swi4, Mbp1
Ime1, Ume6
Msn2, Msn4

cell cycle
sporulation
diauxic shift
DNA damage
stress response
all conditions

cell cycle
YMR016C
YLR183C
YIL131C
SWI4
YDR451C
SWI6
STE12
MBP1
MCM1
YDR146C
YLR131C

sporulation
UME6
IME1
YNL216W
SIN3
YIR023W
YPL038W
YNL103W
YMR021C
CBP1
YBL021C
YIL122W

diauxic shift
HAP4
HAP2
YHR206W
YAP1
HSF1
YPL089C
YCR065W
CINS
YDR310C

DNA damage
YDR259C
MSN2
YDR501W
MSN4
YGL096W
PDR1
YLR403W
YGL071W
YIR018W

stress response
YKL043W
YLR013W
YGL209W
YML027W
YPF034C
YHL009C
YBR049C
YGL035C
YKL112W
YDR043C
YPR065W

Luscombe et al. Nature 431: 308
transient hubs
permanent hubs
Unknown functions

cell cycle
sporulation
diauxic shift
DNA damage
stress response
all conditions

YMR016C
YLR181C
YIL131C
SWI4
YDR451C
SWI6
STE12
MBP1
MCM1
YDR146C
YLR131C

UME6
IME1
YNL216W
SIN3
YIR023W
YPL038W
YNL103W
YMR021C
CBP1
YBL021C
YIL122W

HAP4
YHR206W
YAP1
HSP1
YPL089C
YCR065W
CIN5
YDR310C

YDR259C
MSN2
YDR501W
MSN4
YGL096W
YDR143W
YGL071W
YIR018W

YKL043W
YLR013W
YGL209W
YML027W
YPF034C
YHL009C
YBR049C
YGL035C
YKL112W
YDR043C
YPR065W

Luscombe et al. Nature 431: 308
Network Dynamics #2: Environments

How do molecular networks change across environments?
What pathways are used more?
Used as a biosensor?
What is metagenomics?

Genomics Approach

Culture Microbes → Extract DNA → Sequence

ATCGTATA
CGCGAAG
ACGTCTGA
AGTGTCTGCT

Assemble and Annotate

PROBLEM: Estimated that less than 1% can be cultured in the lab

Metagenomics Approach

Collect Sample → Extract DNA → Sequence

ATCGTGAAGATGAAGATAGA
ATCGTGGCATACGCATCGC
ACAGTGGCTAGCTAAGCTA
CAGCTGACTACCTAGCTAAGCTA
AGTACGCTAGCTAAGCTAAGCTA
AGTACGCTAGCTAAGCTAAGCTA
AGTACGCTAGCTAAGCTAAGCTA
AGTACGCTAGCTAAGCTAAGCTA
AGTACGCTAGCTAAGCTAAGCTA
AGTACGCTAGCTAAGCTAAGCTA
AGTACGCTAGCTAAGCTAAGCTA
AGTACGCTAGCTAAGCTAAGCTA
AGTACGCTAGCTAAGCTAAGCTA
AGTACGCTAGCTAAGCTAAGCTA

PROBLEM: Lose information about which gene belongs to which microbe.
Global Ocean Survey Statistics (GOS)

6.25 GB of data
7.7M Reads
1 million CPU hours to process

Rusch, et al., PLOS Biology 2007
Expressing data as matrices indexed by site, env. var., and pathway usage

[Rusch et. al., (2007) PLOS Biology; Gianoulis et al., PNAS (in press, 2009)]
Simple Relationships: Pairwise Correlations

Gianoulis et al., PNAS (in press, 2009)
Canonical Correlation Analysis: Simultaneous weighting

\[
\text{UPI} = a \times \text{GRE} + b + c \times \text{GPA}
\]

\[
\text{GPI} = a' \times \text{GRE} + b' + c' \times \text{GPA}
\]

[Gianoulis et al., PNAS (in press, 2009)]
Canonical Correlation Analysis: Simultaneous weighting

<table>
<thead>
<tr>
<th>Score</th>
<th># of papers published</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Undergraduate Performance Index (UPI)</th>
<th>Graduate School Performance Index (GPI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE</td>
<td>GPA</td>
</tr>
</tbody>
</table>

**Environmental Features**
- Temp
- Chlorophyll

**Metabolic Pathways**
- Photosynthesis
- Lipid Metabolism

[Gianoulis et al., PNAS (in press, 2009)]
The goal of this technique is to interpret cross-variance matrices. We do this by defining a change of basis.

Given \( X = \{x_1, x_2, \ldots, x_n\} \) and \( Y = \{y_1, y_2, \ldots, y_m\} \)

\[
C = \frac{\sum X \sum_{X,Y}}{\sum Y \sum_{Y,X}}
\]

\[
\max \ Corr(U,V) = \frac{a' \sum_{12} b}{\sqrt{a' \sum_{11} a \sqrt{b' \sum_{22} b}}}
\]

[ Gianoulis et al., PNAS (in press, 2009) ]
DPM: Discriminative Partition Matching

Cluster (Partition)

Environment

Site-Set 1

B1

B3

Site-Set 2

B2

B4

B5

Test

DPM FOOTPRINT

B1

B3

B2

B4

B5

Metabolism

Functional class | pval
--- | ---
InfoStorage & Processing | .07
Cellular Process | .08
Metabolism | 4x10^-14

Taurine biosynthesis
Heme biosynthesis
Asparagine degradation
Nitrogen fixation
Acylglycerol degradation
Asparagine biosynthesis
Cysteine Metabolism

[ Gianoulis et al., PNAS (in press, 2009) ]
Strength of Pathway co-variation with environment

[ Gianoulis et al., PNAS (in press, 2009) ]
Conclusion #1: energy conversion strategy, temp and depth

[ Gianoulis et al., PNAS (in press, 2009) ]
Conclusion #2: Outer Membrane components vary the environment

[ Gianoulis et al., PNAS (in press, 2009) ]
Why is their fluctuation in amino acid metabolism? Is there a feature(s) that underlies those that are environmentally-variant as opposed to those which are not?

[ Gianoulis et al., PNAS (in press, 2009) ]
Conclusion #4: Cofactor (Metal) Optimization

**IS DEPENDENT-ON**

- Methionine synthesis
  - Cobalamin biosynthesis
  - Cobalt transporters

**RELIES ON**

- Methionine Salvage
  - Spermidine/Putrescine transporters
  - Arg/His/Ornithine transporters

**Methionine salvage, synthesis, and uptake, transport**

**IS NEEDED FOR**

- Methionine degradation
  - S-adenosyl Methionine Biosynthesis (synthesize SAM one of the most important methyl donors)
  - Polyamine biosynthesis

[ Gianoulis et al., PNAS (in press, 2009) ]
Biosensors: Beyond Canaries in a Coal Mine

[ Gianoulis et al., PNAS (in press, 2009) ]
Networks & Variation

Which parts of the network vary most in sequence?
Which are under selection, either positive or negative?
METHODOLOGY: MAP SNP AND CNV DATA ONTO ENSEMBL GENES, AND THEN MAP ENSEMBL GENES TO THE KNOWN INTERACTOME


Source: PMK
ADAPTIVE EVOLUTION CAN BE SEEN ON TWO DIFFERENT LEVELS

Intra-species variation

Single-basepair

Positive Selection

Fixed mutations
(differences to other species)

Source: PMK
POSITIVE SELECTION LARGELY TAKES PLACE AT THE NETWORK PERIPHERY

Positive selection in the human interactome

CENTRAL PROTEINS ARE LESS LIKELY TO BE UNDER POSITIVE SELECTION

Peripheral genes are likely to undergo positive selection, whereas hubs aren’t. This is likely due to the following reasons:

- Hubs have stronger structural constraints, the network periphery doesn’t.
- Most recently evolved functions (e.g., ‘environmental interaction genes’ such as sensory perception genes etc.) would probably lie in the network periphery.

Effect is independent of any bias due to gene expression differences.

*With a probability of over 80% to be positively selected as determined by Ka/Ks. Other tests of positive selection (McDonald Kreitmann and LDD) corroborate this result.

CENTRAL NODES ARE LESS LIKELY TO LIE INSIDE OF SDs

This result also confirms our initial hypothesis – peripheral nodes tend to lie in regions rich in SDs.

Since segmental duplications are a different mechanism of ongoing evolution, the less constrained peripheral proteins are enriched in them.

Note that despite the small size of our dataset for known SD’s we get significant correlations. It is to be expected that the correlations will get clearer as more data emerges.*

*Specifically, a number of the SDs are likely not fixed, but rather common CNVs in the reference genome

Why do we observe this? Perhaps central hub proteins are involved in more interactions & have more surface buried.

**BURIED SITES ARE CONSERVED AND MUCH LESS LIKELY TO HARBOR NON-SYNONYMOUS MUTATIONS**

---

**dN/dS Ratio**

- Exposed sites: 0.49
- Buried sites: 0.35

*P* << 0.01

**Average Relative Surface Exposure**

- Site with Synonymous Mutations only: 2.26
- Sites with Non-synonymous Mutations: 2.66

*P* << 0.01

Source: Kim et al. PNAS (2007)
Another explanation: THE NETWORK PERIPHERY CORRESPONDS TO THE CELLULAR PERIPHERY

Source: Gandhi et al. (Nature Genetics 2006), Kim et al. PNAS (2007)
IS RELAXED CONSTRAINT OR ADAPTIVE EVOLUTION THE REASON FOR THE PREVALENCE OF BOTH SELECTED GENES AND SDs AT THE NETWORK PERIPHERY?

<table>
<thead>
<tr>
<th></th>
<th>Relaxed Constraint</th>
<th>Adaptive Evolution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inter-Species Variation (Fixed differences)</strong></td>
<td>• Increases inter-species variation – more variable loci are under less negative selection</td>
<td>• Increases inter-species variation – more variable loci are under less negative selection</td>
</tr>
<tr>
<td></td>
<td>• Can be seen in higher Ka/Ks ratio or SD occurrence</td>
<td>• Can be seen in higher Ka/Ks ratio or SD occurrence</td>
</tr>
<tr>
<td><strong>Intra-Species Variation (Polymorphisms)</strong></td>
<td>• Increases intra-species variation – for the very same reason</td>
<td>• Should not have effects on intra-species variation</td>
</tr>
<tr>
<td></td>
<td>• Can be seen in both SNPs or CNVs</td>
<td></td>
</tr>
</tbody>
</table>

Source: Kim et al. PNAS (2007)
SOME, BUT NOT ALL OF THE SINGLE-BASEPAIR SELECTION AT THE PERIPHERY IS DUE TO RELAXED CONSTRAINT

• There is a difference in variability (in terms of SNPs) between the network periphery and the center

• However, this difference is much smaller than the difference in selection

• This most likely means, that part of the effect we’re seeing is due to relaxed constraint (and higher variability)

• But, not the entire effect*

*But it’s hard to quantify

Source: Kim et al. (2007) PNAS
Similar Results for Large-scale Genomic Changes (CNVs and SDs)

Inter vs. Intra-Species Variation in Networks

Inter-Species (SDs)

Betweenness Centrality (x 10^4)

Genes intersecting SDs

All others

p<0.01

2.61

4.18

Intra-Species (CNVs) [ Variability ]

Betweenness Centrality (x 10^4)

Genes intersecting CNVs

All others

p<0.01

3.25

4.20

Reasoning

• There a small difference in variability (in terms of CNVs) between the network periphery and the center

• But, there is a (as shown before) marked difference in fixed (and hence, presumably, selected) SDs at the network periphery and center

Source: Kim et al. (2007) PNAS
Networks & Variation 2

Which parts of the network vary most in sequence?
More generally which features are most correlated with evolutionary rate
Protein evolutionary rate

Evolutionary rate at the whole protein level

Why do some proteins evolve slowly, while others evolve quickly?

[Slide from Y Xia]
## Many Gene Features Potentially Correlated with Evolutionary Rate

**Meta-features**

<table>
<thead>
<tr>
<th>Meta-features</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino Acid Composition</td>
<td>Amino Acid Content (20 total attributes)</td>
</tr>
<tr>
<td>Structure (Physicochemical Properties)</td>
<td>Predicted helix content</td>
</tr>
<tr>
<td></td>
<td>Predicted sheet content</td>
</tr>
<tr>
<td></td>
<td>Predicted coil content</td>
</tr>
<tr>
<td></td>
<td>Predicted native disorder</td>
</tr>
<tr>
<td></td>
<td>Predicted transmembrane helix content</td>
</tr>
<tr>
<td></td>
<td>Charge (pl)</td>
</tr>
<tr>
<td></td>
<td>Hydrophobicity (Kyte-Doolittle)</td>
</tr>
<tr>
<td></td>
<td>Aromaticity</td>
</tr>
<tr>
<td></td>
<td>Size</td>
</tr>
<tr>
<td>Function</td>
<td>Biological process (GO slim)</td>
</tr>
<tr>
<td></td>
<td>Molecular function (GO slim)</td>
</tr>
<tr>
<td></td>
<td>Cellular compartment (GO slim)</td>
</tr>
<tr>
<td>Abundance</td>
<td>Absolute mRNA expression</td>
</tr>
<tr>
<td></td>
<td>Protein expression</td>
</tr>
<tr>
<td></td>
<td>Codon Adaptation Index (CAI)</td>
</tr>
<tr>
<td></td>
<td>Codon bias</td>
</tr>
<tr>
<td>Phenotype</td>
<td>Essentiality</td>
</tr>
<tr>
<td></td>
<td>Marginal essentiality</td>
</tr>
<tr>
<td>Network</td>
<td>Number of interactors</td>
</tr>
<tr>
<td></td>
<td>Number of transcriptional regulators</td>
</tr>
<tr>
<td>Genome</td>
<td>Degree of gene duplication</td>
</tr>
<tr>
<td></td>
<td>GC content</td>
</tr>
</tbody>
</table>
Assess Relationship of Many Different Features to Evolutionary Rate in Yeast

[Xia et al. ('09). Plos CB]
Simple Relationships between Individual Features & Evolutionary Rate

<table>
<thead>
<tr>
<th>Feature Description</th>
<th>Rank Correlation with Evolutionary Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Codon bias</td>
<td>-0.578</td>
</tr>
<tr>
<td>Codon adaptation index</td>
<td>-0.557</td>
</tr>
<tr>
<td>Protein expression</td>
<td>-0.486</td>
</tr>
<tr>
<td>Absolute mRNA expression</td>
<td>-0.467</td>
</tr>
<tr>
<td>Gly content</td>
<td>-0.401</td>
</tr>
<tr>
<td>Ala content</td>
<td>-0.390</td>
</tr>
<tr>
<td>Ser content</td>
<td>0.366</td>
</tr>
<tr>
<td>Asn content</td>
<td>0.317</td>
</tr>
<tr>
<td>Val content</td>
<td>-0.293</td>
</tr>
<tr>
<td>Native disorder</td>
<td>0.251</td>
</tr>
<tr>
<td>GC content</td>
<td>-0.242</td>
</tr>
<tr>
<td>Degree of gene duplication</td>
<td>-0.206</td>
</tr>
<tr>
<td>Sheet content</td>
<td>-0.191</td>
</tr>
<tr>
<td>Number of interactors</td>
<td>-0.160</td>
</tr>
<tr>
<td>Essentaility</td>
<td>-0.147</td>
</tr>
<tr>
<td>Marginal essentaility</td>
<td>-0.146</td>
</tr>
<tr>
<td># of transcriptional regulators</td>
<td>-0.142</td>
</tr>
<tr>
<td>Hydrophobicity</td>
<td>-0.141</td>
</tr>
<tr>
<td>Leu content</td>
<td>0.105</td>
</tr>
<tr>
<td>Gln content</td>
<td>0.081</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pcbi.1000413.t002

[Xia et al. ('09). Plos CB]
Correlations among top correlates of evolutionary rate

[Xia et al. ('09). Plos CB]
Information contribution of protein features & "meta-features" in the task of predicting slowly evolving proteins

[Xia et al. ('09). Plos CB]
Outline: Molecular Networks

• Why Networks?
• Predicting Networks (yeast)
  ◦ Propagating known information
• Dynamics & Variation of Networks
  ◦ Across cellular states (yeast)
  ◦ Across environments (in prokaryotes)
• Protein Networks & Variation (yeast & humans)
Conclusions on Networks: Predictions

• Predicting Networks
  ◊ Extrapolating from the Training Set
  ◊ Principled ways of using known information in the fullest possible fashion
  • Prediction Propagation
  • Multi-level learning
Conclusions: Network Dynamics across Cellular States

- Merge expression data with Networks
- Active network markedly different in different conditions
- Identify transient hubs associated with particular conditions
- Use these to annotate genes of unknown function
Conclusions: Networks Dynamics across Environments

- Developed and adapted techniques to connect quantitative features of environment to metabolism.
- Applied to available aquatic datasets, we identified footprints that were predictive of their environment (potentially could be used as biosensor).
- Strong correlation exists between a community’s energy conversion strategies and its environmental parameters (e.g. temperature and chlorophyll).
- Suggest that limiting amounts of cofactor can (partially) explain increased import of amino acids in nutrient-limited conditions.
Conclusions: Connecting Networks & Variation

- We find ongoing evolution (positive selection) at the network periphery.
  - This trend is present on two levels:
    - On a sequence level, it can be seen as positive selection of peripheral nodes.
    - On a structural level, it can be seen as the pattern of SDs that display significantly higher allele frequencies in non-central genes.
  - 2 possible mechanisms for this: adaptive evolution at cellular periphery & relaxation of structural constraints at the network periphery.
    - We show that the latter can only explain part of the increased variability.
Conclusions: Connecting Networks & Variation 2

- Evolutionary rate is related to network positioning
- However, only a weak relationship, with more association with abundance and composition
TopNet – an automated web tool

Similar tools include Cytoscape.org, Idekar, Sander et al.

Normal website + Downloaded code (JAVA)
+ Web service (SOAP) with Cytoscape plugin

[Yu et al., NAR (2004); Yip et al. Bioinfo. (2006); Similar tools include Cytoscape.org, Idekar, Sander et al]
Acknowledgements

Networks.GersteinLab.org
Acknowledgements

Networks.GersteinLab.org
Acknowledgements

Networks.GersteinLab.org
More Information on this Talk

**TITLE:** Understanding Protein Function on a Genome-scale through the Analysis of Molecular Networks

**SUBJECT:** Networks

**DESCRIPTION:**
Summit on Systems Biology 2009, The Microbial World and Beyond, Richmond, VA, 2009.05.19, 13:00-14:00; [I:3RDSUMMIT] (Long networks talk, adding in for the first time: evolrate*. Fits easily into 50’ w. 10’ questions. PPT works on mac & PC and has many photos.)

(Paper references in the talk were mostly from Papers.GersteinLab.org. The above topic list can be easily cross-referenced against this website. Each topic abbrev. which is starred is actually a papers “ID” on the site. For instance, the topic pubnet* can be looked up at http://papers.gersteinlab.org/papers/pubnet )

**PERMISSIONS:** This Presentation is copyright Mark Gerstein, Yale University, 2008. Please read permissions statement at http://www.gersteinlab.org/misc/permissions.html. Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).

**PHOTOS & IMAGES**. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info. In particular, many of the images have particular EXIF tags, such as kwpotppt, that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt. 